STAT 8025

Lecture 4: Estimation and Prediction (I)

Dr. Emily Lei Kang

Division of Statistics & Data Science
Department of Mathematical Sciences
University of Cincinnati

University of Cincinnati

Suppose we are interested in a spatial process $\{Y(\mathrm{s}) : \mathrm{s} \in \mathcal{D}\}$. We have data $\mathrm{Y} = (Y(\mathrm{s}_1), \ldots, Y(\mathrm{s}_n))'$ and would like to fit a model and predict $Y(\mathrm{s}_0)$.

- Strategies:
  1. Variogram:
     - Make assumptions (e.g., intrinsic stationarity, weak stationarity)
     - Estimating variogram
     - Kriging (Spatial BLUP)
  2. Maximum likelihood:
     - Make assumptions: GP
     - MLE for parameters
     - Prediction using conditional distribution from multivariate normal distribution
  3. Bayesian inference
     - Make assumptions: GP, priors
     - MCMC for estimation and prediction

Overview of the Variogram Method

1. Using exploratory techniques, prior knowledge, etc., posit a model for $\{Y(\mathrm{s})\}$
2. **Detrend:** Estimate $\boldsymbol{\beta}$ by ordinary least squares, or the mean via median polish, or some other method to estimate the mean that does not require knowledge of the second-order dependence structure.
3. Using fitted residuals (called **detailed residuals**) from the previous step, estimate $\gamma(\mathrm{h})$ (or $C(\mathrm{h})$) nonparametrically, called the **empirical semivariogram** and plot it (in several directions if needed).
4. Select a valid semivariogram model $\gamma(\mathrm{s};\boldsymbol{\theta})$ (or covariance function $C(\mathrm{h};\boldsymbol{\theta})$) that is compatible with the plot

Overview of the Variogram Method, cont'd

5. Fit the chosen model to the empirical semivariogram (or covariance) to estimate the model's parameters via optimization
6. Using the fitted semivariogram or covariance function, re-estimate β by generalized least squares (or by some other method that accounts for second-order dependence structure)
7. "**Krige**" (i.e. predict) unobserved values at sites (or over regions) and estimate the corresponding variances of prediction error.

Detrend

- Why do we need to detrend?
  - We typically need to assume intrinsic stationarity when estimating semivariogram. This implies that we need to assume constant mean. Detrending will help.
- Some ways to detrend:
  - Choose covariates, model $\mu(s) = X(s)'\beta$, and estimate $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$ (same as in regression). $\hat{\beta}_{OLS}$ is not BLUE but can still be unbiased.
  - Using basis function: splines, wavelets, etc.
  - Using locally weighted least squares (LOESS)
    - Assumes that the mean function is smooth
    - Estimates this smooth trend in a moving fashion by fitting a site-specific polynomial.
    - Fits using weighted least squares, with weights inversely related to distance from the site.
  - Median polish

Median Polish

Median polish: The mean function at spatial location (x_l, y_k) is: to be:

$$\mu(x_l, y_k) = a + r_k + c_l.$$

Earlier we saw how median polish fits this model to the data. The fitted values at data locations are:

$$\hat{\mu}(x_l, y_k) = \hat{a} + \hat{r}_k + \hat{c}_l.$$

The fitted surface over the remainder of the spatial domain is obtained by linearly interpolating. For $s_0 = (x, y)$ between the four nodes (x_l, y_k), (x_{l+1}, y_k),(x_l, y_{k+1}),(x_{l+1}, y_{k+1}), the fit is given by the planar interpolant:

$$\hat{a} + \hat{r}_k + \left(\frac{y - y_k}{y_{k+1} - y_k}\right)(\hat{r}_{k+1} - \hat{r}_k) + \hat{c}_l + \left(\frac{x - x_l}{x_{l+1} - x_l}\right)(\hat{c}_{l+1} - \hat{c}_l).$$

(Cressie p. 188 and p. 219)

Variogram Estimation

- It is hard to directly use the sample covariance as we don't have replication
- We need to make some assumptions and then pool information across pairs of locations by distance/difference d in order to estimate
- This is what we do when estimating variogram.
- Recall the definition of semivariogram $\gamma(\mathrm{d})$:

$$2\gamma(\mathrm{d}) = E[(Y(\mathrm{s}) - Y(\mathrm{u}))^2] \text{ where } \mathrm{d} = \mathrm{s} - \mathrm{u}$$

Semivariogram Estimation

- The raw ingredients for semivariogram estimation are either:
  - the observations $\{Y(\mathrm{s}_1), \ldots, Y(\mathrm{s}_n)\}$, if the mean function is taken to be constant;
  - the detailed residuals after detrending

$$\hat{\delta}(\mathrm{s}_i) = Y(\mathrm{s}_i) - \hat{\mu}(\mathrm{s}_i; \hat{\boldsymbol{\beta}}), \quad (i = 1, \ldots, n)$$

    from a fitted mean function at the data locations. E.g., $\hat{\mu}$ with $\hat{\boldsymbol{\beta}}_{OLS}$.
- The basic idea is to estimate $\gamma(\mathrm{h})$ by one-half the average squared difference of responses or residuals whose data locations are lagged by h.

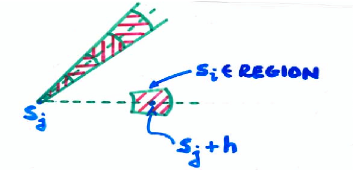
The empirical semivariogram $\hat{\gamma}$ is

$$\hat{\gamma}(\mathrm{h}_u) = \frac{1}{2N(\mathrm{h}_u)} \sum_{\mathcal{B}(\mathrm{h}_u)} (\hat{\delta}(\mathrm{s}_i) - \hat{\delta}(\mathrm{s}_j))^2; u = 1, \ldots, K$$

where $N(\mathrm{h}_u)$ are the number of data pairs s_i and s_j separated by h_u.

- Here $\mathrm{h}_1, \ldots, \mathrm{h}_K$ are the distinct values of h
- $N(\mathrm{h}_u)$ is the number of times that lag h_u occurs in the data set. (We don't double-count.)
- This is a method-of-moments type estimator.
- The estimator is **biased** when the observations themselves are used $(Y(\mathrm{s}_1), \ldots, Y(\mathrm{s}_n))$ and the mean is not a constant. It is approximately unbiased when the detailed residuals are used.
- If the mean is constant $\mu(\mathrm{s}) = \mu$, then the estimator is **unbiased** when the observations themselves are used.

In practice, replace "$s_i - s_j = h$" with "$s_i - s_j \in T(h)$", where $T(h)$ is a *tolerance region* (bin) about h [similar to using histogram to estimate a density]



Then, we obtain

$$2\hat{\gamma}(h_u) \equiv ave\{(\hat{\delta}(s_i) - \hat{\delta}(s_j))^2 : s_i - s_j \in T(h_u)\}$$

where $u = 1, \ldots, K$, and $h_1, \ldots, h_K$ are chosen lags

- Note that we only have estimates of $\gamma(\mathrm{h})$ for a finite number of lags.
- How many lag classes (i.e. how fine a partition) should we use? A rule of thumb is to require $N(\mathrm{h}_u) \geq 30$ and h_u to be less than half the maximum lag represented in the dataset. But there is no harm in trying several different partitions.
- An alternative and more robust (less sensitive to outliers) estimator, proposed by Cressie and Hawkins (1980, *Journal of the International Association for Mathematical Geology*), is

$$\bar{\gamma}(\mathrm{h}_u) = \frac{\{\frac{1}{N(H_u)} \sum_{N(H_u)} |\hat{\delta}(\mathrm{s}_i) - \hat{\delta}(\mathrm{s}_j)|^{1/2}\}^4}{.914 + [.988/N(H_u)]}$$

$$(u = 1, \ldots, K)$$

Covariance Estimation

Recall, for a weakly (second-order) stationary process $Y(\cdot)$,

$$cov(Y(\mathrm{s}), Y(\mathrm{u})) = C(\mathrm{s} - \mathrm{u})$$

We can then estimate the covariance function:

$$\hat{C}(\mathrm{h}) = ave\{(Y(\mathrm{s}_i) - \bar{Y})(Y(\mathrm{s}_j) - \bar{Y}) : \mathrm{s}_i - \mathrm{s}_j = \mathrm{h}\}$$

where $\bar{Y} = \sum Y(\mathrm{s}_i)/n$.

- This estimator is the spatial generalization of the sample autocovariance function used by time series analyses
- This estimator is meaningful only if the process is *weakly (second-order) stationary*; otherwise it's estimating something that doesn't exist.
- When we don't assume constant mean, you can detrend first and use detailed residuals.

Comparison with Semivariogram Estimation

- $\hat{\gamma}(\mathrm{h}) \neq \hat{C}(0) - \hat{C}(\mathrm{h})$, but the difference is usually small for large n.
- It is more common to work with the variogram
  - If the estimates are based on the observations themselves, then $\hat{C}(\mathrm{h})$ is **biased** even when the mean is constant but $\hat{\gamma}(\mathrm{h})$ is **unbiased**.
  - If the estimates are based on detailed residuals from a fitted mean function, then $\hat{\gamma}(\mathrm{h})$ is less biased than $\hat{C}(\mathrm{h})$
  - If there is a trend in the data that is not removed, $\hat{\gamma}(\mathrm{h})$ is not as badly biased as $\hat{C}(\mathrm{h})$. That is, variogram estimator is less sensitive to mean misspecification.

Variogram Model Fitting

- We are not satisfied with the empirical variogram and we don't use the empirical variogram directly when we perform spatial prediction
  - The empirical semivariogram may violate the required property of conditional negative definiteness.
  - For various purposes (e.g. kriging) we may require an estimate of the semivariogram at a lag not represented in the data.
  - The empirical semivariogram may be quite bumpy. A smoothed version may be helpful for understanding the nature of the spatial dependence.
- The empirical variogram can be visualized to suggest an appropriate model.

- Let $\gamma(\mathrm{h}; \boldsymbol{\theta})$ denote the parametric model to be fit to the empirical semivariogram and let Θ denote the parameter space for $\boldsymbol{\theta}$.
- Methods of Fitting
  - By eye...
  - Ordinary nonlinear least squares
  - Weighted nonlinear least squares (Cressie, 1985, Mathematical Geology)
  - Generalized least squares? Derivation and calculation of $\widehat{\mathrm{var}}(\hat{\gamma})$ can be a challenge.
  - Maximum likelihood

One weighted nonlinear least squares estimator of $\gamma(\mathrm{h};\boldsymbol{\theta})$ is defined as a value $\hat{\boldsymbol{\theta}} \in \Theta$ that minimizes the weighted residual sum of squares function:

$$\sum_{u=1}^{K} \frac{N(\mathrm{h}_u)}{[\gamma(\mathrm{h}_u;\boldsymbol{\theta})]^2}[\hat{\gamma}(\mathrm{h}_u) - \gamma(\mathrm{h}_u;\boldsymbol{\theta})]^2.$$

- The weights $\frac{N(\mathrm{h}_u)}{[\gamma(\mathrm{h}_u;\boldsymbol{\theta})]^2}$ are small if either $N(\mathrm{h}_u)$ is small or $\gamma(\mathrm{h}_u;\boldsymbol{\theta})$ is large. Thus, nonparametric estimates at large lags tend to receive relatively less weight.

Revisit of Mean Estimation

- When we detrend, we didn't account for the spatial dependence. For example, assuming $\mu(\mathrm{s}) = \mathrm{X}(\mathrm{s})'\boldsymbol{\beta}$, we use

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathrm{X}'\mathrm{X})^{-1}(\mathrm{X}'\mathrm{Y})$$

  where $\mathrm{Y} = (Y(\mathrm{s}_1), \ldots, Y(\mathrm{s}_n))'$

- Assume $E(\mathrm{Y}) = \mathrm{X}\boldsymbol{\beta}$ and $Var(\mathrm{Y}) = \Sigma$ with the (ij)-th element $C(\mathrm{s}_i, \mathrm{s}_j)$. Is our OLS estimator $\hat{\beta}_{OLS}$ good?
  - It is unbiased but not the 'best' among all unbiased linear estimators.
  - It is better to use the generalized least squares estimator:

$$\hat{\boldsymbol{\beta}}_{GLS} = [\mathrm{X}'\Sigma(\boldsymbol{\theta})^{-1}\mathrm{X}]^{-1}(\mathrm{X}'\Sigma(\boldsymbol{\theta})^{-1}\mathrm{Y})$$

- A common approach is to plug-in $\hat{\theta}$, giving estimated $\hat{\boldsymbol{\beta}}_{GLS}$ (EGLS).

$$\hat{\boldsymbol{\beta}}_{EGLS} = [\mathrm{X}'\Sigma(\hat{\boldsymbol{\theta}})^{-1}\mathrm{X}]^{-1}(\mathrm{X}'\Sigma(\hat{\boldsymbol{\theta}})^{-1}\mathrm{Y})$$

$$\hat{\beta}_{EGLS} = [X'\Sigma(\hat{\theta})^{-1}X]^{-1}(X'\Sigma(\hat{\theta})^{-1}Y)$$

- Is this OK?
  - We can *approximately* quantify the uncertainty of $\hat{\beta}_{EGLS}$ via

$$\hat{\mathrm{var}}(\hat{\beta}_{EGLS}) = (X'\Sigma(\hat{\theta})^{-1}X)^{-1}$$

although it tends to underestimate $\mathrm{var}(\hat{\beta}_{EGLS})$

Kriging (Ch. 3 of SSD)

$$\text{kriging} \equiv \text{(spatial) BLUP}$$

The origin of the word *kriging* is from D.G. Krige, a South African mining engineer who in the 1950's developed empirical methods for predicting ore grades at unsampled locations using the known grades of ore sampled at nearby sites.
[See "Origins of kriging" by Cressie, 1990]

Write $Y(\mathrm{s}) = \mu(\mathrm{s}) + \delta(\mathrm{s})$; $\mathrm{s} \in \mathcal{D}$
where $E(\delta(\mathrm{s})) = 0$ and $E(Y(\mathrm{s})) = \mu(\mathrm{s})$

$$cov(Y(\mathrm{s}), Y(\mathrm{u})) = E(\delta(\mathrm{s}), \delta(\mathrm{u})) = C(\mathrm{s}, \mathrm{u})$$

or

$$E(\delta(\mathrm{s}) - \delta(\mathrm{u}))^2 = 2\gamma(\mathrm{s}, \mathrm{u})$$

with

- $\mu(\mathrm{s})$: mean function
- $C(\mathrm{s}, \mathrm{u})$: covariance function
- $2\gamma(\mathrm{s}, \mathrm{u})$: variogram function

Predict $Y(\mathbf{s}_0)$ from data $\mathbf{Y} \equiv (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))'$ where $\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_n$ are known.

- For now, we assume that there is no measurement error. Later, we will assume that we observe $Z(\mathbf{s}) = Y(\mathbf{s}) + \epsilon(\mathbf{s})$ where $\epsilon(\mathbf{s})$ represents measurement error, and we would like to predict $Y(\mathbf{s}_0)$ from noisy data $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))'$.
- We will begin with the assumption that $\mu(\mathbf{s}) = \mu$ but μ is unknown. The resulting kriging is called **ordinary kriging**.

Ordinary Kriging

$$Y(\mathsf{s}) = \mu + \delta(\mathsf{s}); \mathsf{s} \in \mathcal{D}$$

- We consider only predictors that are *linear and unbiased*:

$$\hat{Y}(\mathsf{s}_0) = \sum_i \lambda_i Y(\mathsf{s}_i) + k$$

$$E(\hat{Y}(\mathsf{s}_0)) = E(Y(\mathsf{s}_0)) = \mu, \text{ for all } \mu \in \mathbb{R}$$

Thus,

$$\sum_i \lambda_i = 1; k = 0$$

- We would like to find the optimal prediction that minimizes the squared error loss. That is to find $\{\lambda_i\}$ s.t. $\sum_i \lambda_i = 1$ and

$$MSPE \equiv E(Y(\mathsf{s}_0) - \sum \lambda_i Y(\mathsf{s}_i))^2 \text{ is minimized.}$$

OK in terms of $\gamma(\cdot)$: MSPE of Linear Unbiased Predictor

Algebraic result:
Given

$$\sum_i a_i = 0 = \sum_j b_j$$

we have

$$\{\sum_i a_i Z(\mathrm{s}_{1i})\}\{\sum_j b_j Z(\mathrm{s}_{2j})\} = -(1/2)\sum_i \sum_j a_i b_j \{Z(\mathrm{s}_{1i}) - Z(\mathrm{s}_{2j})\}^2$$

MSPE of a Linear Unbiased Predictor $\hat{Y}(s_0) = \sum_{i=1}^n \lambda_i Y(s_i)$

- $\hat{Y}(s_0) = \sum_{i=1}^n \lambda_i Y(s_i)$ is unbiased. So we have $\sum_{i=1}^n \lambda_i = 1$.
- Use the result on the previous slide:

$$\begin{aligned} MSPE &= E(Y(s_0) - \sum \lambda_i Y(s_i))^2 \\ &= -\sum_{i=0}^n \sum_{j=0}^n a_i a_j \gamma(s_i, s_j) \end{aligned}$$

where $a_0 = 1$, $a_i = -\lambda_i$ and notice that $\sum_{i=0}^n a_i = 0$.

- Therefore,

$$\begin{aligned} MSPE(\boldsymbol{\lambda}) &= -\sum_{i=0}^n \sum_{j=0}^n a_i a_j \gamma(s_i, s_j) \\ &= 2\sum_{i=1}^n \lambda_i \gamma(s_0, s_i) - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(s_i, s_j) \end{aligned}$$

since $\gamma(s_i, s_i) = 0$.

Ordinary Kriging, cont'd

- Thus, with $\boldsymbol{\lambda} \equiv (\lambda_1, \ldots, \lambda_n)'$ and $\boldsymbol{\lambda}'1 = 1$, we would like to minimize $MSPE(\boldsymbol{\lambda})$
- Using Lagrange multiplier m, we minimize:

$$MSPE(\boldsymbol{\lambda}) - 2m(\sum_{j=1}^{n} \lambda_j - 1),$$

w.r.t. $\boldsymbol{\lambda}$, and Lagrange multiplier m.

Solve

$$\frac{\partial}{\partial \lambda_i}\{MSPE(\boldsymbol{\lambda}) - 2m(\sum_{j=1}^{n} \lambda_j - 1)\} = 0; i = 1, \ldots, n$$

$$\sum_{j=1}^{n} \lambda_j - 1 = 0$$

That is, solve:

$$2\gamma(s_0, s_i) - 2\sum_{j=1}^{n} \lambda_j \gamma(s_i, s_j) - 2m = 0$$

$$\sum_{j=1}^{n} \lambda_j - 1 = 0$$

That is,

$$\Gamma_O \boldsymbol{\lambda}_O = \gamma_O$$

Ordinary Kriging Equations (pp. 119-127 of SSD)

$$\boldsymbol{\lambda}_O = \Gamma_O^{-1}\boldsymbol{\gamma}_O$$

$$\left[\begin{array}{c} \lambda_{O,1} \\ \vdots \\ \lambda_{O,n} \\ \hline m \end{array}\right] = \left[\begin{array}{ccc|c} & & & 1 \\ & \gamma(s_i, s_j) & & \vdots \\ & & & 1 \\ \hline 1 & \cdots & 1 & 0 \end{array}\right]^{-1} \left[\begin{array}{c} \gamma(s_0, s_1) \\ \vdots \\ \gamma(s_0, s_n) \\ \hline 1 \end{array}\right]$$

Kriging variance:

$$
\begin{aligned}
\sigma_k^2(\mathsf{s}_0) &= E(Y(\mathsf{s}_0) - \textstyle\sum_{i=1}^n \lambda_{O,i} Y(\mathsf{s}_i))^2 \\
&= 2\textstyle\sum_{i=1}^n \lambda_i \gamma(\mathsf{s}_0, \mathsf{s}_i) - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathsf{s}_i, \mathsf{s}_j) \\
&= 2\textstyle\sum_{i=1}^n \lambda_i \gamma(\mathsf{s}_0, \mathsf{s}_i) - \sum_{i=1}^n \{\sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathsf{s}_i, \mathsf{s}_j)\} \\
&= 2\textstyle\sum_{i=1}^n \lambda_i \gamma(\mathsf{s}_0, \mathsf{s}_i) - \sum_{i=1}^n \lambda_i \{\gamma(\mathsf{s}_i, \mathsf{s}_0) - m\} \\
&= \boldsymbol{\lambda}_O' \gamma_O = \gamma_O' \Gamma_O^{-1} \gamma_O
\end{aligned}
$$

Ordinary Kriging Equations

Then we can derive the OK formulas:

$$\hat{Y}_{OK}(s_0) = \left\{ \gamma + 1 \frac{1 - 1'\Gamma^{-1}\gamma}{1'\Gamma^{-1}1} \right\}' \Gamma^{-1} Y$$

$$\sigma_k^2(s_0) = \gamma'\Gamma^{-1}\gamma - \frac{(1 - 1'\Gamma^{-1}\gamma)^2}{1'\Gamma^{-1}1}$$

- From the ordinary kriging formulas (*without M.E.*), show that if $s_0 = s_i$, then

$$\hat{Y}_{OK}(s_i) = Y(s_i)$$

Proof: WLOG, assume $s_0 = s_1$, then γ_O equals the first column of Γ_O.
Since $\Gamma_O^{-1}\Gamma_O = I$,

$$\boldsymbol{\lambda}_O = \Gamma_O^{-1}\gamma_O = (1, 0, \cdots, 0)'$$

Thus, $\hat{Y}_{OK}(s_i) = Y(s_i)$.

Remarks on Ordinary Kriging

- Ordinary kriging is derived under the assumption of constant mean.
  - Kriging in Practice
    1. Detrend (the trend is not necessarily linear)
    2. Perform OK using the detailed residuals
    3. Prediction = trend + OK predictor
- OK is derived under the assumption that the semivariogram is known. In practice, the semivariogram is unknown and must be estimated, and the estimated $\hat{\gamma}(\cdot)$ replaces $\gamma(\cdot)$ in the kriging equations and in the expression for the kriging variance.
  - The estimated kriging variance tends to underestimate the prediction error variance of the OK predictor because it does not account for the estimation error incurred in estimating θ, parameters in semivariogram.

- Ok is a linear combination of *all* the observations. In practice, sometimes only the observations within a *moving window* or *kriging neighborhood* are used; local kriging.
- *Environmental monitoring programs.* Note that the kriging variance at any given site s_0 does not depend on the data. thus, it can be used to answer sampling design questions, such as where to take one more observation to maximize the reduction in σ^2_{OK} at a certain point, or where to take one more observation to minimize the maximum (or average) value of σ^2_{OK} over the entire spatial domain. Same idea is used in *computer model calibration*.

Measurement Error (pp. 127-130 of SSD)

Assume additive measurement errors:

$$Z(\mathrm{s}) = Y(\mathrm{s}) + \epsilon(s); \mathrm{s} \in \mathcal{D}$$

where $\epsilon(\cdot)$ is zero-mean white noise, independent of $Y(\cdot)$, and $var(\epsilon(\mathrm{s})) = \tau^2 > 0$
We want to predict $Y(\mathrm{s}_0)$, from data $\mathrm{Z} \equiv (Z(\mathrm{s}_1), \ldots, Z(\mathrm{s}_n))'$ with a linear predictor:

$$\hat{Y}(\mathrm{s}_0) = \sum_{i=1} \lambda_i Z(\mathrm{s}_i) + k$$

OK with M.E.

We assume constant mean $\mu(\cdot) = \mu$. For $\hat{Y}(s_0) = \sum_i \lambda_i Z(s_i) + k$, we require:
Uniform unbiasedness:

$$E(\hat{Y}(s_0)) = E(\sum_i \lambda_i Z(s_i) + k) = \mu; \text{ for all } \mu \in \mathbb{R}$$

Thus, we have

$$\sum_i \lambda_i = 1; k = 0$$

Spatial Best Linear Unbiased Predicion (BLUP):
Find $\{\lambda_i\}$ s.t. $\sum_i \lambda_i = 1$ and

$$MSPE(\boldsymbol{\lambda}) \equiv E(Y(s_0) - \sum_i \lambda_i Z(s_i))^2 \text{ is minimized}$$

This is ordinary kriging (OK) in its most general form (includes measurement error)

OK with ME

The optimal $\boldsymbol{\lambda}_O = (\lambda_1, \ldots, \lambda_n, m)'$ is given by

$$\boldsymbol{\lambda}_O^* = \Gamma_O^{-1} \boldsymbol{\gamma}_O^*,$$

where

$$\boldsymbol{\gamma}_O^* \equiv (\gamma^*(\mathrm{s}_0, \mathrm{s}_1), \ldots, \gamma^*(\mathrm{s}_0, \mathrm{s}_n), 1)'$$

$$\gamma^*(\mathrm{s}, \mathrm{u}) = \begin{cases} \tau^2; & \mathrm{s} = \mathrm{u} \\ \gamma(\mathrm{s}, \mathrm{u}); & \mathrm{s} \neq \mathrm{u} \end{cases}$$

and recall

$$2\gamma(\mathrm{s}, \mathrm{u}) \equiv var(Z(\mathrm{s}) - Z(\mathrm{u}))$$

The minimized MSPE (kriging variance) is

$$\sigma_k^2(\mathrm{s}_0) \equiv E(Y(\mathrm{s}_0) - \sum_i \lambda_{O,i}^* Z(\mathrm{s}_i))^2 = \boldsymbol{\gamma}_O^{*\prime} \Gamma_O^{-1} \boldsymbol{\gamma}_O^* - \tau^2$$

- Derivation of OK with ME is similar as OK (skipped in class)
- WLOG let $\mathsf{s}_0 = \mathsf{s}_1$ then $\hat{Y}(\mathsf{s}_0) \neq Z(\mathsf{s}_1)$

OK in Terms of the Covariance Function (p. 123 of SSD)

Recall OK: Find $\{\lambda_i\}$ for $\hat{Y}(\mathrm{s}_0) \equiv \sum \lambda_i Y(\mathrm{s}_i)$ s.t. $\sum \lambda_i = 1$ (uniform unbiasedness) and $MSPE(\boldsymbol{\lambda}) \equiv E(Y(\mathrm{s}_0) - \sum \lambda_i Y(\mathrm{s}_i))^2$ is minimized.

Minimize:

$$\begin{aligned} MSPE(\boldsymbol{\lambda}) - 2m(\textstyle\sum \lambda_i - 1) = \; & C(\mathrm{s}_0, \mathrm{s}_0) + \textstyle\sum_i \sum_j \lambda_i \lambda_j C(\mathrm{s}_i, \mathrm{s}_j) \\ & - 2 \textstyle\sum_i \lambda_i C(\mathrm{s}_0, \mathrm{s}_i) - 2m(\textstyle\sum \lambda_j - 1) \\ = \; & C(\mathrm{s}_0, \mathrm{s}_0) + \boldsymbol{\lambda}' \Sigma \boldsymbol{\lambda} \\ & - 2\boldsymbol{\lambda}' \mathrm{c}(\mathrm{s}_0) - 2m(1'\boldsymbol{\lambda} - 1) \end{aligned}$$

Then we get:

$$\begin{bmatrix} \lambda_{O,1} \\ \vdots \\ \lambda_{O,n} \\ \hline -m \end{bmatrix} = \left[\begin{array}{c|c} & 1 \\ C(\mathrm{s}_i, \mathrm{s}_j) & \vdots \\ & 1 \\ \hline 1 \quad \cdots \quad 1 & 0 \end{array}\right]^{-1} \begin{bmatrix} C(\mathrm{s}_0, \mathrm{s}_1) \\ \vdots \\ C(\mathrm{s}_0, \mathrm{s}_n) \\ \hline 1 \end{bmatrix}$$

i.e., $(\lambda_{O,1}, \cdots, \lambda_{O,n}, -m)' = \mathrm{C}_O^{-1}\mathrm{c}_O$

Ordinary Kriging Equations

Then we can derive the OK formulas:

$$\hat{Y}_{OK}(\mathrm{s}_0) = \left\{ \mathrm{c}(\mathrm{s}_0) + 1 \frac{1 - 1'\Sigma^{-1}\mathrm{c}(\mathrm{s}_0)}{1'\Sigma^{-1}1} \right\}' \Sigma^{-1}\mathrm{Y}$$

$$\sigma^2_{OK}(\mathrm{s}_0) = C(\mathrm{s}_0, \mathrm{s}_0) - \mathrm{c}(\mathrm{s}_0)'\Sigma^{-1}\mathrm{c}(\mathrm{s}_0) + \frac{(1 - 1'\Sigma^{-1}\mathrm{c}(\mathrm{s}_0))^2}{1'\Sigma^{-1}1}$$

Universal Kriging (pp. 151-183 of SSD)

Model:

$$Y(\mathrm{s}) = \sum_{j=0}^{p} \beta_j x_j(\mathrm{s}) + \varepsilon(\mathrm{s}) \equiv \mathrm{x}(\mathrm{s})\boldsymbol{\beta}' + \delta(\mathrm{s})$$

where $\delta(\cdot)$ is a zero-mean geostatistical process with variogram $\gamma(\cdot,\cdot)$ (or covariance function $C(\cdot,\cdot)$)

Predictor:

$$\hat{Y}(\mathrm{s}_0) = \sum_{i=1}^{n} \lambda_i Y(\mathrm{s}_i)$$

Uniformly Unbiased:

$$E(\sum \lambda_i Y(\mathrm{s}_i)) = E(Y(\mathrm{s}_0)), \text{ for all } \boldsymbol{\beta} \in \mathbb{R}^{p+1}$$

Spatial BLUP:

Minimize

$$MSPE(\boldsymbol{\lambda}) \equiv E(Y(\mathrm{s}_0) - \sum \lambda_i Y(\mathrm{s}_i))^2,$$

subject to the uniform unbiasedness constraint

Universal Kriging, cont'd

MSPE:

$$MSPE(\boldsymbol{\lambda}) = 2\sum \lambda_i \gamma(\mathrm{s}_i, \mathrm{s}_0) - \sum\sum \lambda_i \lambda_j \gamma(\mathrm{s}_i, \mathrm{s}_j)$$

Uniform Unbiasedness:

$$E(Y(\mathrm{s}_0)) = E(\sum \lambda_i Y(\mathrm{s}_i))$$

$$\implies \mathrm{x}(\mathrm{s}_0)'\boldsymbol{\beta} = \boldsymbol{\lambda}'\mathrm{X}\boldsymbol{\beta}, \text{ for all } \boldsymbol{\beta}$$

$$\implies \mathrm{x}(\mathrm{s}_0)' = \boldsymbol{\lambda}'\mathrm{X},$$

where X is the $n \times (p+1)$ matrix and $\mathrm{x}(\mathrm{s}_0) \equiv (x_0(\mathrm{s}_0), \ldots, x_p(\mathrm{s}_0))'$.

Minimize:

$$MSPE(\boldsymbol{\lambda}) \text{ subject to } \mathrm{X}'\boldsymbol{\lambda} = \mathrm{x}(\mathrm{s}_0)$$

i.e., minimize:

$$MSPE(\boldsymbol{\lambda}) - 2\mathrm{m}'(\mathrm{X}'\boldsymbol{\lambda} - \mathrm{x}(\mathrm{s}_0)),$$

w.r.t. $\boldsymbol{\lambda}$ and Lagrange multipliers $\mathrm{m} \equiv (m_0, \ldots, m_p)'$.

Universal Kriging Equations

Solve

$$\Gamma_U \boldsymbol{\lambda}_U = \boldsymbol{\gamma}_U$$

$$\begin{bmatrix} \lambda_{U,1} \\ \vdots \\ \lambda_{U,n} \\ \hline m_0 \\ \vdots \\ m_p \end{bmatrix} \begin{bmatrix} & & & x_0(s_1) & \cdots & x_p(s_1) \\ & \gamma(s_i, s_j) & & \vdots & & \vdots \\ & & & x_0(s_n) & \cdots & x_p(s_n) \\ \hline x_0(s_1) & \cdots & x_0(s_n) & & & \\ \vdots & & \vdots & & 0 & \\ x_p(s_1) & \cdots & x_p(s_n) & & & \end{bmatrix}^{-1} \begin{bmatrix} \gamma(s_0, s_1) \\ \vdots \\ \gamma(s_0, s_n) \\ \hline x_0(s_0) \\ \vdots \\ x_p(s_0) \end{bmatrix}$$

Then the UK formulas can be derived (details skipped):

$$\hat{Y}_{UK}(\mathrm{s}_0) = \left\{\gamma(\mathrm{s}_0) + \mathrm{X}(\mathrm{X}'\Gamma^{-1}\mathrm{X})^{-1}(\mathrm{x}(\mathrm{s}_0) - \mathrm{X}'\Gamma^{-1}\gamma(\mathrm{s}_0))\right\}' \Gamma^{-1}\mathrm{Y}$$

$$\sigma^2_{UK}(\mathrm{s}_0) = \gamma(\mathrm{s}_0)'\Gamma^{-1}\gamma(\mathrm{s}_0) - (\mathrm{x}(\mathrm{s}_0) - \mathrm{X}'\Gamma^{-1}\gamma(\mathrm{s}_0))'(\mathrm{X}'\Gamma^{-1}\mathrm{X})^{-1}(\mathrm{x}(\mathrm{s}_0) - \mathrm{X}'\Gamma^{-1}\gamma(\mathrm{s}_0))$$

We can also derive the formulas in terms of covariance functions (details skipped):

$$\hat{Y}_{UK}(\mathrm{s}_0) = \left\{\mathrm{c}(\mathrm{s}_0) + \mathrm{X}(\mathrm{X}'\Sigma^{-1}\mathrm{X})^{-1}(\mathrm{x}(\mathrm{s}_0) - \mathrm{X}'\Sigma^{-1}\mathrm{c}(\mathrm{s}_0))\right\}' \Sigma^{-1}\mathrm{Y}$$

$$\begin{aligned}\sigma^2_{UK}(\mathrm{s}_0) = \quad & C(\mathrm{s}_0, \mathrm{s}_0) - \mathrm{c}(\mathrm{s}_0)'\Sigma^{-1}\mathrm{c}(\mathrm{s}_0) \\ & + (\mathrm{x}(\mathrm{s}_0) - \mathrm{X}'\Sigma^{-1}\mathrm{c}(\mathrm{s}_0))'(\mathrm{X}'\Sigma^{-1}\mathrm{X})^{-1}(\mathrm{x}(\mathrm{s}_0) - \mathrm{X}'\Sigma^{-1}\mathrm{c}(\mathrm{s}_0))\end{aligned}$$

Optimal Linear Prediction (pp. 172-177 of SSD)

Model:

$$Y(\mathrm{s}) = \mathrm{x}(\mathrm{s})'\boldsymbol{\beta} + \delta(\mathrm{s}); \mathrm{s} \in \mathcal{D}$$

Data:

$$\mathrm{Y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\delta}; E(\boldsymbol{\delta}) = 0, \mathrm{var}(\boldsymbol{\delta}) = \Sigma$$

Suppose we know $\boldsymbol{\beta}$. To minimize the squared error loss, our best linear predictor is the given by the simple kriging predictor as follows:

$$Y^*(\mathrm{s}_0) = \mathrm{x}(\mathrm{s}_0)'\boldsymbol{\beta} + \mathrm{c}(\mathrm{s}_0)'\Sigma^{-1}(\mathrm{Y} - \mathrm{X}\boldsymbol{\beta}),$$

But β is unknown. What is the best linear unbiased estimator (BLUE) for β?

$$\hat{\beta} \equiv (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$$

It can be shown that

$$\hat{Y}_{UK}(s_0) = \{c(s_0) + X(X'\Sigma^{-1}X)^{-1}(x(s_0) - X'\Sigma^{-1}c(s_0))\}'\Sigma^{-1}Y$$

$$\implies \hat{Y}_{UK}(s_0) = x(s_0)'\hat{\beta} + c(s_0)'\Sigma^{-1}(Y - X\hat{\beta})$$

That is,

$$\text{universal kriging} \equiv \text{best linear prediction} + \text{BLUE of } \beta$$

Summary

- Variogram estimation
- Kriging

Preview:

- MLE and Bayesian inference
- Analysis of large spatial data